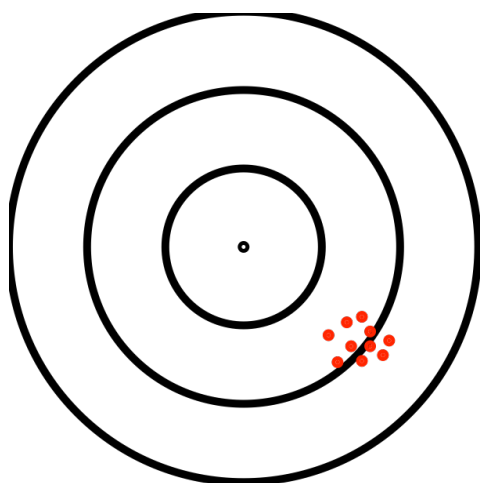


Bias

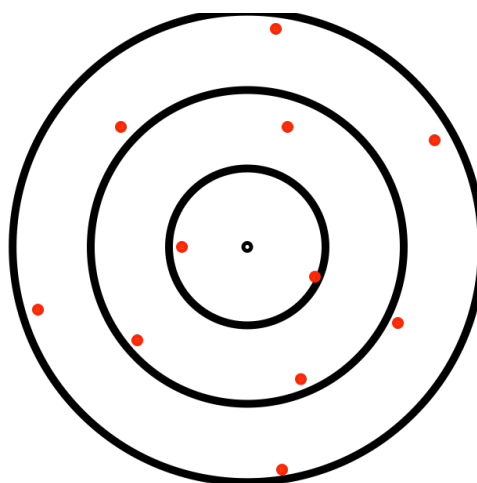
Definition: Any systematic error in the design, conduct or analysis of a study that results in a mistaken estimate of an exposure's effect on the risk of the disease

-Epidemiology (Leon Gordis)

Systematic Error vs Random Error



Systematic Error



Random Error

Types of Bias: there are 3 main types of biases.

1. Selection bias
2. Information bias/ measurement bias
3. Bias due to confounding

Types of bias	
Selection Bias	Measurement bias/ Information bias
Self selection bias	Recall bias
Berkson's Bias	Detection bias
Survivorship (Neyman's) bias	Observer's bias
Healthy worker effect	
Exposure-related bias	

Loss to follow up bias	
Inappropriate control group	

SELECTION BIAS: Selection bias is a systematic error resulting from the way subjects are either selected in a study or else are selectively lost to follow up.

For example: Suppose in a study of asbestos exposure and lung cancer the exposure is distributed among the cases and controls in the target population as follows:

	Diseased	Non diseased
Exposed	100	200
Unexposed	100	400

The true OR in the target population is $(100 \times 400) / (100 \times 200) = 2.0$.
If the selection probabilities for all the cells in table were equal at 90% the 2*2 table of selection probabilities would look like the following:

	Diseased	Non diseased
Exposed	100×0.90	200×0.90
Unexposed	100×0.90	400×0.90

$$OR = (90 \times 360) / (90 \times 180) = 2.0$$

Here we have an example with each cell in our study population containing the same proportion of subjects as the corresponding cell in the target population. 90% of each cell is sampled. In this case, selection bias does not exist.

If the selection probabilities are unequal, but still proportional, we still do not observe any selection bias in our study. If the selection probability is 90% among the diseased individuals and the selection probability is 70% among the non-diseased individuals the resulting 2x2 table would look like the following.

	Diseased	Non diseased
Exposed	$100 \times 0.90 = 90$	$200 \times 0.70 = 140$
Unexposed	$100 \times 0.90 = 90$	$400 \times 0.70 = 280$

$$OR = (90 \times 280) / (90 \times 140)$$

If however the selection probabilities are unequal, and also non-proportional, then selection bias will occur. The following table shows how selection bias occurs when the selection probability for the unexposed controls is different than that of the other three groups of study members

	Diseased	Non diseased
Exposed	$100 \times 0.90 = 90$	$200 \times 0.90 = 180$
Unexposed	$100 \times 0.90 = 90$	$400 \times 0.70 = 280$

$$OR = (90 \times 280) / (90 \times 180) = 1.6$$

To accurately represent the target population, we need the selection odds for exposure among the diseased (α/β) to be equal to the selection odds for exposure among the non-diseased (γ/δ). If the selection odds are different, then selection bias will distort our study measure of effect from the "truth", which in this study is 2.0. Likewise, no selection bias will occur if the selection odds for disease among the exposed is equal to the selection odds for disease among the nonexposed, i.e. if $\alpha/\gamma = \beta/\delta$. The two previous statements can be combined into one general principle—No selection bias will occur if the cross product of the four selection probabilities is equal to one, i.e. $(\alpha \times \delta) / (\beta \times \gamma) = 1$. Selection bias will occur in cohort studies if the rates of participation or the rates of loss to follow-up differ by both exposure and disease status. Although we seldom can know the exposure and disease status of nonrespondents or persons lost to follow-up, it is sometimes possible to obtain these data from an external source.

Self selection Bias (volunteers induced bias): individuals who volunteer for a study possess different characteristics than the average individual in the target population.

Berksons' bias (hospital selective admission): This can be a problem in case-control studies. It occurs because patients with two concurrent diseases or health problems are more likely to be admitted to a hospital than those with a single condition.

For example, people who have both peptic ulcers and also who smoke are more likely to be admitted to a hospital than people who have either of them. A case control study trying to evaluate the relationship between smoking and peptic ulcers may therefore find a much stronger association between the two than would really exist in general community.

Incidence-Prevalence bias (Survivorship bias, Neyman's bias):

This is a major issue in case-control and cross-sectional studies. A bias that occurs when we try to estimate the risk of a disease on the basis of data collected at a given time point in a series of survivors rather than on data gathered during a certain time period in a group of incident cases. It arises when a gap in time occurs between exposure and selection of study participants

For example a case control study to evaluate the protective effect of physical exercise on MI was undertaken by taking cases of MI and healthy controls and asking them about the history of regular physical exercise. Surprisingly, a large number of both the cases and controls give a history of regular physical exercise; the study concluded that regular physical exercise does not protect against MI. we know that 25% to 33% of the cases of acute MI die within the first 3 hrs. Only those who live get admitted to the hospital and are available as cases. Now regular physical exercise may be an important factor in helping the person to overcome the acute myocardial episode. Thus out of the cases of MI, those who did not undertake regular exercise died, while the ones who did exercise were the ones who lived to give such a history.

Healthy Worker effect: A comparison between health status of military and civilian population may show a better health status of the soldiers; one of the important reasons may be because of the initial medical examination during which the unfit persons are excluded and only 'healthy workers' are included in the army. The basic dictum of selection and comparisons in research should be to 'compare likes with likes'

Exposure related bias: this is a special type of Berkson's bias. If the hospital admission probability is different among those who have and those who do not have the suspected cause, such a selection bias can occur. This is specially liable to occur in case control studies. For example, such an exposure related selection bias was viewed with concern in a case-control study that found an association between use of dietary supplementation with L-Tryptophan and 'Eosinophilia- Myalgia syndrome'. The main criticism was that the initial press publicity about a suspected association may have resulted in a preferential diagnosis among known users of L-Tryptophan as compared to non-users. Thus the estimate of risk (OR) obtained from such studies may have overestimated the true effect of risk.

Bias due to loss to follow up: this is a special problem in cohort and experimental studies. If the subjects drop out/are withdrawn/ die before assessment of outcome/ do not respond later on/ cross over to follow-up could have been systematically different from those who continued.

Bias due to selection of inappropriate control: this is another major issue in case-control studies. The major dictum that should be followed in case-control study is that

controls should be derived from same source population from which cases have come and the controls should be equally at risk. For example in case- control study which desired to assess the risk associated with non-use of condoms (exposure) with development of STD (outcome). Investigator selected cases from a STD clinic and also controls from same STD clinic who were found to be free of STD after evaluation, at this clinic. However, many of these may not have developed STD probably because they had sex partner who himself/herself had STD, and hence subjects had no chance of exposure whether they used condom or not. Hence right choice of control group in this research would have been to take people who were known sex partners of persons known to be having STD but were found clear for STD, while cases should have been those who had known STD persons as sex partners and were detected to be having STD.

Information Bias/ Measurement Bias:

Information bias can occur when the means for obtaining information about the subjects in the study are inadequate so that as a result some of the information gathered regarding exposures and/or disease outcome is incorrect.

Misclassification bias: given inaccuracies in methods of data collection, may at sometimes misclassify subjects and introduce misclassification bias. This is of 2 types:

Differential misclassification bias: the rate of misclassification differs in different study groups. For example, misclassification of exposure may occur such that cases are misclassified as being exposed more often than controls are. Example women who had baby with malformation tended to remember more mild infections that occurred during their pregnancies than did mothers of normal infant. Thus there was tendency of differential misclassification in regard to prenatal infection, in that more unexposed cases were misclassified as exposed than were unexposed controls. The result was an apparent association of malformations with infections even though nonexistent. So a differential misclassification bias can lead to either to an apparent association even if one does not really exist or to an apparent lack of association when one does in fact exist.

Nondifferential misclassification bias: it results from degree of inaccuracy that characterizes how information is obtained from any study group- either cases or controls or exposed and non-exposed persons. Such misclassification is not related to exposure status or to a case or control status it is just a problem inherent in the data collection methods. The amount and direction of misclassification is same in cases and controls. The usual effect of non-differential misclassification is that the relative risk or odds ratio tends to be diluted, and it is shifted towards 1.0. We are less likely to detect an association even if one really exists. For example everyone may underreport their own or their spouse's habitual alcohol consumption.

Types of information bias:

1. **Recall bias**: This is a major problem in case - control as also in cross - sectional studies. The fact that a person has become diseased, he or she is more likely to recall the possible exposure; e.g. in a study of X-ray exposure during pregnancy and subsequent leukaemia in children, mothers of leukaemic children are likely to recall more and thus give more history of X-ray exposures.
2. **Reporting bias**: reporting bias occur when a subject feels reluctant to report the history of exposure. E.g. In a study to observe the relationship of induced abortion to risk of breast cancer, reporting bias might played a role in those case control studies that reported a positive association: healthy control may have been more reluctant than women with breast cancer to report that they had an induced abortion.
3. **Observer's (interviewer's bias)**: If the interviewer is aware as to which group is having the particular exposure (in a follow up study) or the disease (in a case control study) then he/she would be more inclined (subconsciously) to interrogate/ examine that particular group more exhaustively, to prove the research question.
4. **Surveillance bias**: if a population is monitored over a period of time, disease ascertainment may be better in monitored population than in general population and may introduce bias which leads to erroneous estimate of relative risk or odds ratio
5. **Wish bias**: term wish bias was coined by Wynder and co workers to denote the bias introduced by subjects who have developed a disease and who in

attempting to answer the question 'why me', seek to show often unintentionally that the disease is not their fault. Thus they may deny certain exposures related to lifestyle (such as smoking and drinking).

Control of selection bias:

There are 4 main strategies for control of selection bias

1. Sampling the cases and controls in the same way
2. Matching
3. Using two or more controls
4. Using a population-based sample

Control of measurement error

Measurement error can be minimized by

- Development of explicit, objective criteria for measuring environmental characteristics and health outcomes
- Careful consistent data collection- for example, through use of standardized instruments; objectives, closed ended questionnaires; valid instruments
- Careful consistent use of data instruments- for example, through use of standardized training and instruction manuals, blinding to the extent possible
- Development and application of quality control/ quality assurance procedures
- Use of multiple sources of data
- Data cleaning and coding

Analysis and adjustment, if necessary, to take account of measurement bias

Biases in Cohort study:

There are five broad categories of bias that are operative in cohort studies.

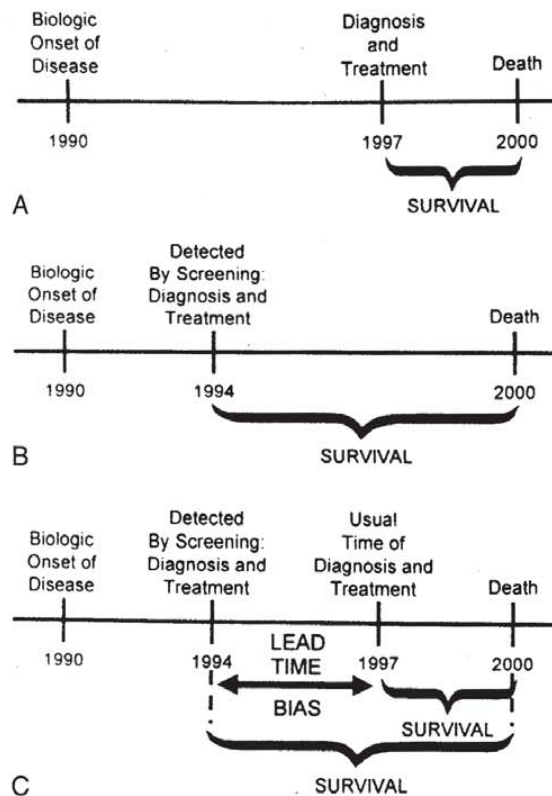
1. Selection bias
2. Follow-up bias
3. Information bias
4. Bias due to confounding
5. Post hoc bias.

Biases in screening programmes

- 1. Volunteers bias**
- 2. Lead time bias**
- 3. Length time bias**
- 4. Overdiagnosis bias**

Lead time bias: figure A shows the natural history of the disease in a hypothetical patient with colon cancer, which was diagnosed in the usual clinical context without any screening. Biological onset of the disease was in 1990. The patient became aware of symptoms in 1997, and had a diagnostic workup leading to a diagnosis of colon cancer. Surgery was performed in 1997 and the patient died of colon cancer in 2000. This patient has survived for 3 years. If we use 5 year survival as an index of treatment success, this patient is a treatment failure.

Consider what might happen to this patient if he resides in a community in which a screening program is initiated. For this hypothetical example only, let us assume that there is actually no benefit from early detection- that is, the natural history of colon cancer is unaffected by early intervention. In this case the patient is asymptomatic but undergoes a routine screening test in 1994, the result of which is positive. In 1994 surgery is performed and the patient died in 2000. The patient has survived 6 years and now clearly a 5 year survivor. However, he is a 5 year survivor not because death has been delayed, but because diagnosis has been made earlier. When we compare this scenario with the scenario without screening, it is apparent the patient has not derived any benefit from earlier detection in terms of having lived any longer; indeed the patient may have lost out in terms of quality of life, as the earlier detection of disease by screening gave him an additional 3 years of postoperative and other him of 3 years of normal life. This problem of illusion of better survival only because earlier detection is called the lead time bias.



Length time biases: Length time bias is a form of selection bias, a statistical distortion of results, which can lead to incorrect conclusions about the data. Length time bias can occur when the lengths of intervals are analyzed by selecting intervals that occupy randomly chosen points in time or space. This process favors longer intervals, thus skewing the data.

Length time bias is often discussed in the context of the benefits of cancer screening, where it can lead to the perception that screening leads to better outcomes when in reality it has no effect. Fast-growing tumors generally have a shorter asymptomatic phase than slower-growing tumors. This means that there is a shorter period of time when the cancer is present in the body (and therefore might be detected by screening) but not yet large enough to cause symptoms, which would cause the patient to seek medical care and be diagnosed without screening. As a result, if the same number of slow-growing and fast-growing tumors appear in a year, the screening test will detect more slow-growers than fast-growers. If these slow growing tumors are less likely to be fatal than the fast growers are, the people whose cancer is detected by screening will do better, on average, than the people whose tumors are detected from symptoms (or at autopsy), even if there is no real benefit to catching the cancer earlier. This can give the impression that detecting cancers through screening causes cancers to be less dangerous, when the reality is that less dangerous cancers are simply more likely to be detected by screening

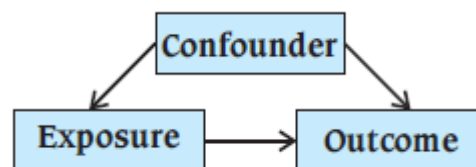
Overdiagnosis bias: Another potential bias is of overdiagnosis. At times, persons who initiate screening program have almost unlimited enthusiasm for the program. Even cytologists reading pap smears may become so enthusiastic that they may tend to overread the smears (false positive readings). Consequently the abnormal group

will be diluted with women who are free of disease. If normal individuals in the screened group are more likely to be diagnosed as positive than are normal individuals in the unscreened group (eg identified as having cancer when in reality they do not), one could get a false impression of increased rates of detection and diagnosis of early-stage disease as a result of screening. In addition, because many of persons with diagnosis of cancer would actually not have cancer, and would therefore have a good survival, the results would represent an inflated estimate of survival after screening in persons thought to have cancer. It is therefore essential that the diagnostic process be rigorously standardized in such studies.

Confounding: A confounding variable is defined as one which explains away the observed association between an exposure and an outcome variable.

For a variable to be a confounder,

- It should be a known risk factor for the disease or the outcome
- It should be associated with the exposure
- It should not be in direct chain or linked between the exposure and outcome
- It should be differentially distributed in the two group.



Example : Age may be a confounding factor in an association between high blood pressure and coronary heart disease (CHD).

This is because, age is a known risk factor for CHD and age is associated with hypertension.

Hypothetical example: unmatched case control study to evaluate the association between HTN and Coronary heart disease.

Table 1: number of exposed and non exposed cases and control

Exposed (HTN)	Cases (CHD+)	Control (CHD -)
Yes	30	18
No	70	82
Total	100	100

$$\text{Odds ratio} = 30 \times 82 / 18 \times 70 = 1.95$$

The question arises that is this association a causal one or could it have resulted from differences in age distribution.

Table 2: distribution of case and control by age

Age (yr)	Cases (CHD+)	Controls (CHD-)
<40	50	80
≥40	50	20
Total	100	100

80% of the control are younger than 40 years of age .50% of the cases are old age as compared to 20% of the control.

Table 3: Relationship of exposure to age

Age	Total	Exposed (HTN+)	Not exposed (HTN-)	%exposed (% HTN+)
<40	130	13	117	10
≥40	70	35	35	50

Among those who were younger than 40 years , (10%) were exposed. Among those who were more than 40 years, 50% were exposed. Hence age is related to HTN or exposure.

Table 4: calculation of OR after stratifying by age

Age yrs	Exposed (HTN)	Case (CHD+)	Control (CHD-)	Odds ratio
<40	Yes	5	8	$5 \times 72 / 45 \times 8 = 360 / 360 = 1.0$
	No	45	72	
	Total	50	80	
≥40	Yes	25	10	$25 \times 10 / 25 \times 10 = 250 / 250 = 1.0$
	No	25	10	
	Total	50	20	

After stratification, odds ratio is 1 in each stratum. Thus the odds ratio of 1.95 was because there was difference in age distributions between those who have CHD and those who do not have CHD. Hence age is a confounder.

Control of confounding:

Several methods are available to control confounding, either through study design or during the analysis of results.

The methods commonly used to control confounding in the design of an epidemiological study are:

- Randomization
 - Restriction
 - Matching
- At the analysis stage, confounding can be controlled by:
- Stratification
 - Statistical modeling or multivariate analysis

Randomization: Randomization, which is applicable only to experimental studies, is the ideal method for ensuring that potential confounding variables are equally distributed among the groups being compared.

Restriction: Restriction can be used to limit the study to people who have particular characteristics. For example, in a study on the effects of coffee on coronary heart disease, participation in the study could be restricted to nonsmokers, thus removing any potential effect of confounding by cigarette smoking.

Matching: Matching is used to control confounding by selecting study participants so as to ensure that potential confounding variables are evenly distributed in the two groups being compared. For example, in a case-control study of exercise and coronary heart disease, each patient with heart disease can be matched with a control of the same age group and sex to ensure that confounding by age and sex does not occur. Matching has been used extensively in case-control studies but it can lead to problems in the selection of controls if the matching criteria are too strict or too numerous; this is called overmatching.

Matching can be expensive and time-consuming, but is particularly useful if the danger exists of there being no overlap between cases and controls, such as in a situation where the cases are likely to be older than the controls.

Stratification: In large studies it is usually preferable to control for confounding in the analytical phase rather than in the design phase. Confounding can then be controlled by stratification, which involves the measurement of the strength of associations in well-defined and homogeneous categories (strata) of the confounding variable

Statistical modeling: Although stratification is conceptually simple and relatively easy to carry out, it is often limited by the size of the study and it cannot help to control many factors simultaneously, as is often necessary. In this situation, multivariate statistical modeling is required to estimate the strength of the associations while controlling for several confounding variables simultaneously.